Project NYPD Shooting Incident Data

Simon F.

2024-06-23

List of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year.

This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included.

The last Metadata updated date was April 26th 2024

Step 0: Import Library

```
# install.packages("tidyverse")
library(tidyverse)
library(lubridate)
```

Step 1: Load Data

- leveraging 'read_csv()' to read the file

```
df = read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")
```

```
## Rows: 28562 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(df) # this shows the headers
```

```
## # A tibble: 6 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
```

```
##          <dbl> <chr>      <time>     <chr>     <chr>                <dbl>
## 1    244608249 05/05/2022 00:10      MANHATTAN INSIDE                  14
## 2    247542571 07/04/2022 22:20      BRONX     OUTSIDE                 48
## 3     84967535 05/27/2012 19:35      QUEENS    <NA>                   103
## 4    202853370 09/24/2019 21:00      BRONX     <NA>                    42
## 5     27078636 02/25/2007 21:00      BROOKLYN  <NA>                    83
## 6    230311078 07/01/2021 23:07      MANHATTAN <NA>                    23
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>,
## #   Longitude <dbl>, Lon_Lat <chr>
```

Step 2: Tidy and Transform Data

Need to get rid of some of the columns, will create new df as df_2 that only has the columsn that I want.

```
df_2 = df %>% select(INCIDENT_KEY,OCCUR_DATE, OCCUR_TIME, BORO, STATISTICAL_MURDER_FLAG, PERP_AGE_GROUP
```

Adding in some unknowns for some missing data

```
df_2 = df_2 %>%
     replace_na(list(PERP_AGE_GROUP = "Unknown", PERP_SEX = "Unknown", PERP_RACE = "Unknown"))
```

Tidying some of the values

```
df_2$PERP_AGE_GROUP = recode(df_2$PERP_AGE_GROUP, UNKNOWN = "Unknown")
df_2$PERP_SEX = recode(df_2$PERP_SEX, U = "Unknown")
df_2$PERP_RACE = recode(df_2$PERP_RACE, UNKNOWN = "Unknown")
df_2$VIC_SEX   = recode(df_2$VIC_SEX, U = "Unknown")
df_2$VIC_RACE   = recode(df_2$VIC_RACE, UNKNOWN = "Unknown")
```

Manking INCIDENT_KEY a character

```
df_2$INCIDENT_KEY = as.character(df_2$INCIDENT_KEY)
```

Coverting the vector into factors

```
df_2$BORO = as.factor(df_2$BORO)
df_2$PERP_AGE_GROUP = as.factor(df_2$PERP_AGE_GROUP)
df_2$PERP_SEX = as.factor(df_2$PERP_SEX)
df_2$PERP_RACE = as.factor(df_2$PERP_RACE)
df_2$VIC_AGE_GROUP = as.factor(df_2$VIC_AGE_GROUP)
df_2$VIC_SEX = as.factor(df_2$VIC_SEX)
df_2$VIC_RACE = as.factor(df_2$VIC_RACE)
```

Run Summary Stats

```
summary(df_2)
```

```
##  INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME                  BORO
##  Length:28562       Length:28562       Length:28562      BRONX        : 8376
##  Class :character   Class :character   Class1:hms        BROOKLYN     :11346
##  Mode  :character   Mode  :character   Class2:difftime   MANHATTAN    : 3762
##                                        Mode  :numeric    QUEENS       : 4271
##                                                          STATEN ISLAND:  807
##
##
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP     PERP_SEX             PERP_RACE
##  Mode :logical           Unknown:12492   (null) : 1141   BLACK         :11903
##  FALSE:23036             18-24  : 6438   F      :  444   Unknown       :11147
##  TRUE :5526              25-44  : 6041   M      :16168   WHITE HISPANIC: 2510
##                          <18    : 1682   Unknown:10809   BLACK HISPANIC: 1392
##                          (null) : 1141                   (null)        : 1141
##                          45-64  :  699                   WHITE         :  298
##                          (Other):   69                   (Other)       :  171
##  VIC_AGE_GROUP      VIC_SEX                                VIC_RACE
##  <18    : 2954   F      : 2760   AMERICAN INDIAN/ALASKAN NATIVE:   11
##  1022   :    1   M      :25790   ASIAN / PACIFIC ISLANDER      :  440
##  18-24  :10384   Unknown:   12   BLACK                         :20235
##  25-44  :12973                  BLACK HISPANIC                : 2795
##  45-64  : 1981                  Unknown                       :   70
##  65+    :  205                  WHITE                         :  728
##  UNKNOWN:   64                  WHITE HISPANIC                : 4283
```

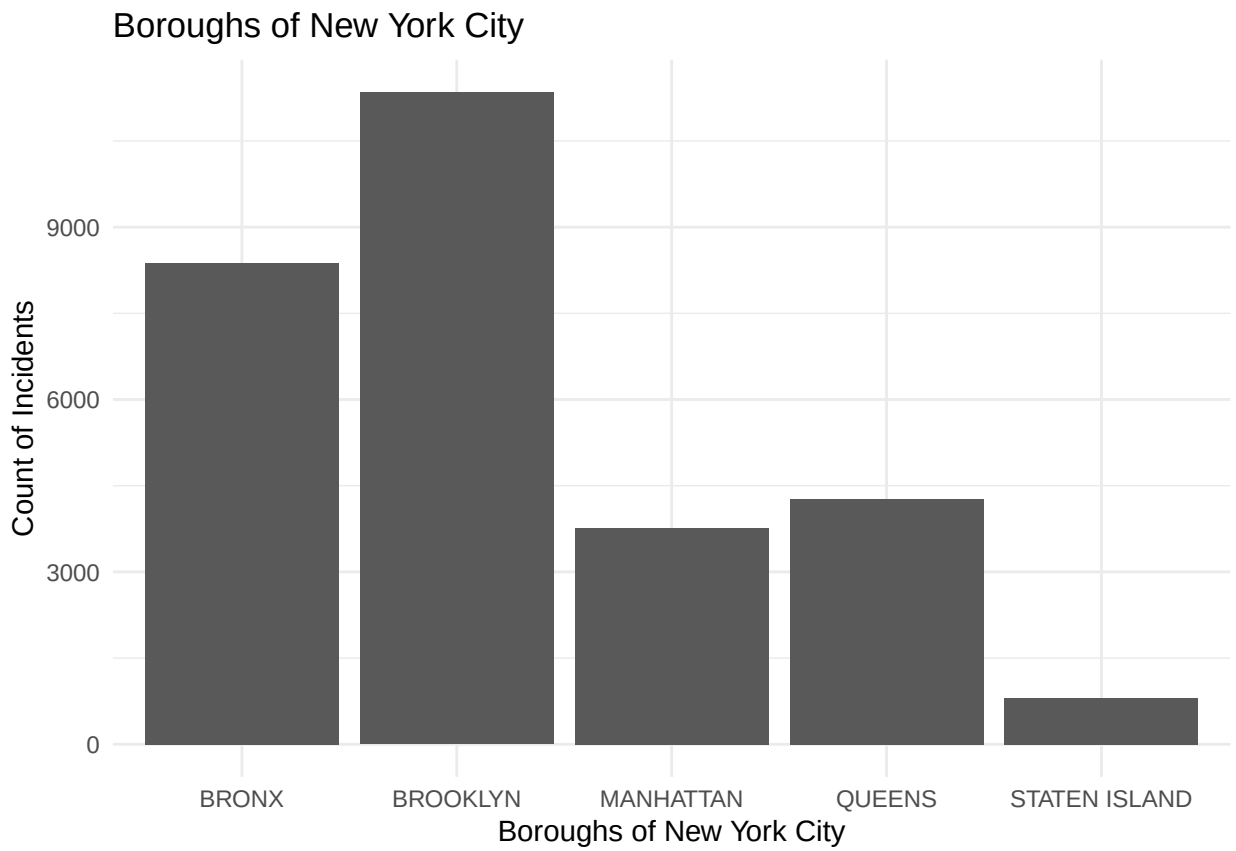
Step 3: Add Visualizations and Analysis

Need to include a few visualizations and one model. 1st visualization will be of which boroughs have the highest counts of incidents

This code shows a bar graph of each boro by its total count if incidents

```
vis_1 <- ggplot(df_2, aes(x = BORO)) +
  geom_bar() +
  labs(title = "Boroughs of New York City",
       x = "Boroughs of New York City",
       y = "Count of Incidents") +
  theme_minimal()

vis_1
```



* We can see that Brooklyn has had the most incidents from all of NYC boroughs.This is followed by Bronx, Queens, Manhattan, then Staten Island.

The 2nd visualization that I want to view is if the incident was a murder or not in the form of a table.

```
table(df_2$BORO, df_2$STATISTICAL_MURDER_FLAG)
```

```
##
##                  FALSE TRUE
##   BRONX           6742 1634
##   BROOKLYN        9136 2210
##   MANHATTAN       3090  672
##   QUEENS          3431  840
##   STATEN ISLAND    637  170
```

The 3rd visualization will be of which time of day incidents occurred.

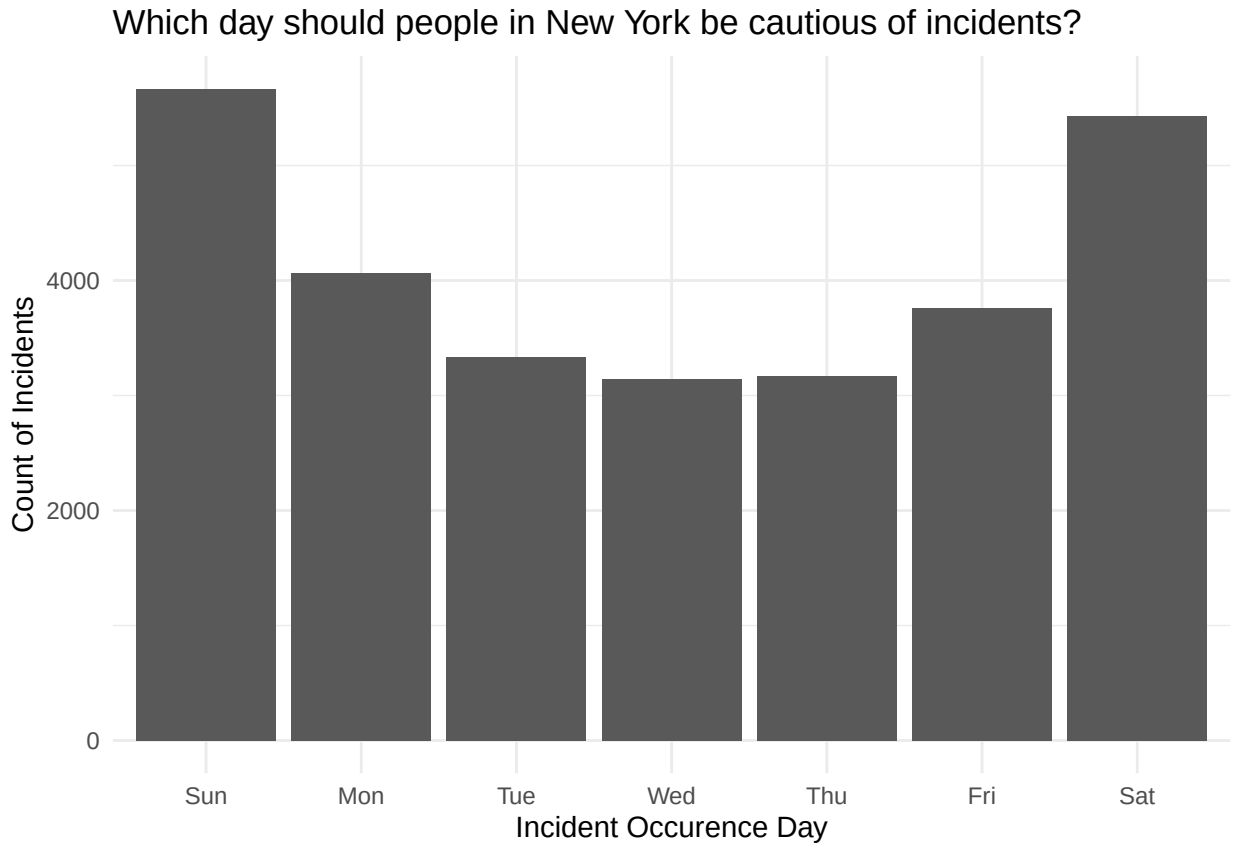
```
df_2$OCCUR_DAY = mdy(df_2$OCCUR_DATE)
df_2$OCCUR_DAY = wday(df_2$OCCUR_DAY, label = TRUE)
df_2$OCCUR_HOUR = hour(hms(as.character(df_2$OCCUR_TIME)))

df_3 = df_2 %>%
  group_by(OCCUR_DAY) %>%
  count()

df_4 = df_2 %>%
```

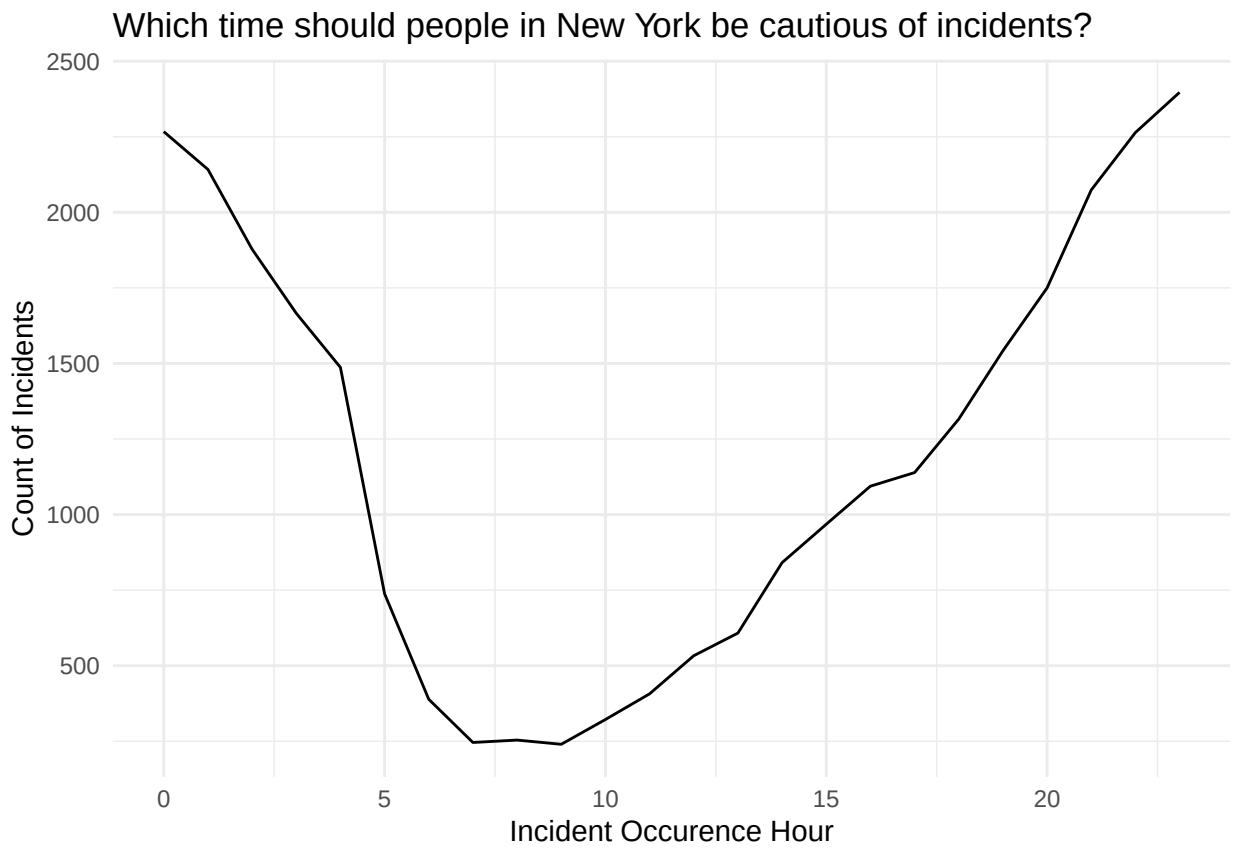
```
  group_by(OCCUR_HOUR) %>%
  count()

vis_3 <- ggplot(df_3, aes(x = OCCUR_DAY, y = n)) +
  geom_col() +
  labs(title = "Which day should people in New York be cautious of incidents?",
       x = "Incident Occurence Day",
       y = "Count of Incidents") +
  theme_minimal()
vis_3
```



* The 4th visualization represents time of day incidents occurred.

```
vis_4 <- ggplot(df_4, aes(x = OCCUR_HOUR, y = n)) +
  geom_line() +
  labs(title = "Which time should people in New York be cautious of incidents?",
       x = "Incident Occurence Hour",
       y = "Count of Incidents") +
  theme_minimal()
vis_4
```



Now, lest look at the Perpetrators and Victims of this data

Age group

```
table(df_2$PERP_AGE_GROUP, df_2$VIC_AGE_GROUP)
```

```
##
##            <18 1022 18-24 25-44 45-64  65+ UNKNOWN
##   (null)   106    0   311   619    96    9       0
##   <18      521    0   652   413    79   15       2
##   1020       0    0     0     1     0    0       0
##   1028       0    0     0     1     0    0       0
##   18-24    808    1  2841  2394   335   47      12
##   224        0    0     1     0     0    0       0
##   25-44    270    0  1560  3600   524   49      38
##   45-64     21    0    85   373   202   13       5
##   65+        0    0     2    27    24   12       0
##   940        0    0     0     1     0    0       0
##   Unknown 1228    0  4932  5544   721   60       7
```

Sex

```
table(df_2$PERP_SEX, df_2$VIC_SEX)
```

```
##
##               F     M Unknown
```

```
##   (null)    123  1018       0
##   F          77   366       1
##   M        1755 14406       7
##   Unknown   805 10000       4
```

Race/Ethnicity

```
table(df_2$PERP_RACE, df_2$VIC_RACE)
```

```
##
##                                 AMERICAN INDIAN/ALASKAN NATIVE
##   (null)                                                    1
##   AMERICAN INDIAN/ALASKAN NATIVE                            0
##   ASIAN / PACIFIC ISLANDER                                  0
##   BLACK                                                     4
##   BLACK HISPANIC                                            0
##   Unknown                                                   5
##   WHITE                                                     0
##   WHITE HISPANIC                                            1
##
##                                 ASIAN / PACIFIC ISLANDER BLACK BLACK HISPANIC
##   (null)                                              27   795            115
##   AMERICAN INDIAN/ALASKAN NATIVE                       0     2              0
##   ASIAN / PACIFIC ISLANDER                            61    56             14
##   BLACK                                              164  9411            839
##   BLACK HISPANIC                                      20   561            365
##   Unknown                                            113  8523            999
##   WHITE                                               13    42             23
##   WHITE HISPANIC                                      42   845            440
##
##                                 Unknown WHITE WHITE HISPANIC
##   (null)                              1    20            182
##   AMERICAN INDIAN/ALASKAN NATIVE      0     0              0
##   ASIAN / PACIFIC ISLANDER            0    12             26
##   BLACK                              25   205           1255
##   BLACK HISPANIC                      6    36            404
##   Unknown                            25   187           1295
##   WHITE                               1   165             54
##   WHITE HISPANIC                     12   103           1067
```

Buidling a model

I want to see view the probability if an incident is likely a murder case or not?

```
mylogit <- glm(STATISTICAL_MURDER_FLAG ~ PERP_RACE + PERP_SEX + PERP_AGE_GROUP + OCCUR_HOUR + OCCUR_DAY
summary(mylogit)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ PERP_RACE + PERP_SEX +
##     PERP_AGE_GROUP + OCCUR_HOUR + OCCUR_DAY, family = binomial,
##     data = df_2)
```

```
##
## Coefficients: (2 not defined because of singularities)
##                                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           -1.746038   0.087418 -19.974  < 2e-16
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE -9.846453 139.210365  -0.071 0.943612
## PERP_RACEASIAN / PACIFIC ISLANDER      1.019563   0.281830   3.618 0.000297
## PERP_RACEBLACK                         0.623191   0.225100   2.769 0.005631
## PERP_RACEBLACK HISPANIC                0.530074   0.234151   2.264 0.023586
## PERP_RACEUnknown                       0.140252   0.087721   1.599 0.109853
## PERP_RACEWHITE                         1.157223   0.256305   4.515 6.33e-06
## PERP_RACEWHITE HISPANIC                0.763092   0.229286   3.328 0.000874
## PERP_SEXF                             -2.455728   0.263527  -9.319  < 2e-16
## PERP_SEXM                             -2.626348   0.239262 -10.977  < 2e-16
## PERP_SEXUnknown                             NA         NA      NA       NA
## PERP_AGE_GROUP<18                      2.217039   0.169750  13.061  < 2e-16
## PERP_AGE_GROUP1020                    -7.816789 196.967745  -0.040 0.968344
## PERP_AGE_GROUP1028                    -7.807111 196.967747  -0.040 0.968383
## PERP_AGE_GROUP18-24                    2.412144   0.160240  15.053  < 2e-16
## PERP_AGE_GROUP224                     -7.872611 196.967750  -0.040 0.968118
## PERP_AGE_GROUP25-44                    2.718384   0.160171  16.972  < 2e-16
## PERP_AGE_GROUP45-64                    3.084714   0.176912  17.436  < 2e-16
## PERP_AGE_GROUP65+                      3.126006   0.303561  10.298  < 2e-16
## PERP_AGE_GROUP940                     -7.883466 196.967753  -0.040 0.968074
## PERP_AGE_GROUPUnknown                       NA         NA      NA       NA
## OCCUR_HOUR                            -0.001020   0.001873  -0.545 0.585949
## OCCUR_DAY.L                           -0.039280   0.037583  -1.045 0.295951
## OCCUR_DAY.Q                           -0.063457   0.040244  -1.577 0.114840
## OCCUR_DAY.C                           -0.053024   0.040579  -1.307 0.191319
## OCCUR_DAY^4                           -0.006062   0.041281  -0.147 0.883256
## OCCUR_DAY^5                            0.026557   0.043402   0.612 0.540619
## OCCUR_DAY^6                           -0.090093   0.044567  -2.022 0.043225
##
## (Intercept)                           ***
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE
## PERP_RACEASIAN / PACIFIC ISLANDER     ***
## PERP_RACEBLACK                        **
## PERP_RACEBLACK HISPANIC               *
## PERP_RACEUnknown
## PERP_RACEWHITE                        ***
## PERP_RACEWHITE HISPANIC               ***
## PERP_SEXF                             ***
## PERP_SEXM                             ***
## PERP_SEXUnknown
## PERP_AGE_GROUP<18                     ***
## PERP_AGE_GROUP1020
## PERP_AGE_GROUP1028
## PERP_AGE_GROUP18-24                   ***
## PERP_AGE_GROUP224
## PERP_AGE_GROUP25-44                   ***
## PERP_AGE_GROUP45-64                   ***
## PERP_AGE_GROUP65+                     ***
## PERP_AGE_GROUP940
## PERP_AGE_GROUPUnknown
## OCCUR_HOUR
```

```
## OCCUR_DAY.L
## OCCUR_DAY.Q
## OCCUR_DAY.C
## OCCUR_DAY^4
## OCCUR_DAY^5
## OCCUR_DAY^6                              *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28061  on 28561  degrees of freedom
## Residual deviance: 27097  on 28536  degrees of freedom
## AIC: 27149
##
## Number of Fisher Scoring iterations: 10
```

Conclusion

My analysis was to see what characteristics, if any, had any impact on if an incident were more likely a murder case or not. I have never really looked into these types of statistics for NYC specifically, but the data does show alot of interesting things. For example, the perp age group really does not have anything to do with if an incident was a murder case or not. The day of the week does tho however, on Thursdays there is more likely to be an incident.

More males commit incidents than females. PACIFIC ISLANDER commit the least amount of incidents *The hour of day does have an impact if an incident is a murder case or not

Thank you